made going beyond the LSDA by adding gradient terms or higher derivatives ($\nabla\rho$ and $\nabla^2\rho$) of the electron density to the exchange–correlation energy or its corresponding potential. In this context several physical constraints can be formulated, which an exact theory should obey. Most approximations, however, satisfy only part of them. For example, the exchange density (needed in the construction of these two quantities) should integrate to -1 according to the Fermi exclusion principle (Fermi hole). Such considerations led to the generalized gradient approximation (GGA), which exists in various parameterizations, *e.g.* in the one by Perdew *et al.* (1996). This is an active field of research and thus new functionals are being developed and their accuracy tested in various applications.

The Coulomb potential $V_C(\mathbf{r})$ in (2.2.10.5) is that of all N electrons. That is, any electron is also moving in its own field, which is physically unrealistic but may be mathematically convenient. Within the HF method (and related schemes) this self-interaction is cancelled exactly by an equivalent term in the exchange interaction [see (2.2.9.1)]. For the currently used approximate density functionals, the self-interaction cancellation is not complete and thus an error remains that may be significant, at least for states (*e.g.* 4*f* or 5*f*) for which the respective orbital is not delocalized. Note that delocalized states have a negligibly small self-interaction. This problem has led to the proposal of self-interaction corrections (SICs), which remove most of this error and have impacts on both the single-particle eigenvalues and the total energy (Parr *et al.*, 1978).

The Hohenberg–Kohn theorems state that the total energy (of the ground state) is a functional of the density, but the introduction of the KS orbitals (describing quasi-electrons) are only a tool in arriving at this density and consequently the total energy. Rigorously, the Kohn–Sham orbitals are *not electronic orbitals* and the KS eigenvalues ε_i (which correspond to $E_{\mathbf{k}}$ in a solids) are *not* directly related to electronic *excitation energies*. From a formal (mathematical) point of view, the ε_i are just Lagrange multipliers without a physical meaning.

Nevertheless, it is often a good approximation (and common practice) to partly ignore these formal inconsistencies and use the orbitals and their energies in discussing electronic properties. The gross features of the eigenvalue sequence depend only to a smaller extent on the details of the potential, whether it is orbital-based as in the HF method or density-based as in DFT. In this sense, the eigenvalues are mainly determined by orthogonality conditions and by the strong nuclear potential, common to DFT and the HF method.

In processes in which one removes (ionization) or adds (electron affinity) an electron, one compares the N electron system with one with $N-1$ or $N+1$ electrons. Here another conceptual difference occurs between the HF method and DFT. In the HF method one may use Koopmans' theorem, which states that the $\varepsilon_i^{\rm HF}$ agree with the ionization energies from state i assuming that the corresponding orbitals do not change in the ionization process. In DFT, the ε_i can be interpreted according to Janak's theorem (Janak, 1978) as the partial derivative with respect to the occupation number n_i,

$$\varepsilon_i = \frac{\partial E}{\partial n_i}. \qquad (2.2.10.8)$$

Thus in the HF method ε_i is the total energy difference for $\Delta n = 1$, in contrast to DFT where a differential change in the occupation number defines ε_i, the proper quantity for describing metallic systems. It has been proven that for the exact density functional the eigenvalue of the highest occupied orbital is the first ionization potential (Perdew & Levy, 1983). Roughly, one can state that the further an orbital energy is away from the highest occupied state, the poorer becomes the approximation to use ε_i as excitation energy. For core energies the deviation can be significant, but one may use Slater's transition state (Slater, 1974), in which half an electron is removed from the corresponding orbital, and then use the $\varepsilon_i^{\rm TS}$ to represent the ionization from that orbital.

Another excitation from the valence to the conduction band is given by the energy gap, separating the occupied from the unoccupied single-particle levels. It is well known that the gap is not given well by taking $\Delta\varepsilon_i$ as excitation energy. Current DFT methods significantly underestimate the gap (half the experimental value), whereas the HF method usually overestimates gaps (by a factor of about two). A trivial solution, applying the 'scissor operator', is to shift the DFT bands to agree with the experimental gap. An improved but much more elaborate approach for obtaining electronic excitation energies within DFT is the GW method in which quasi-particle energies are calculated (Hybertsen & Louie, 1984; Godby *et al.*, 1986; Perdew, 1986). This scheme is based on calculating the dielectric matrix, which contains information on the response of the system to an external perturbation, such as the excitation of an electron.

In some cases, one can rely on the total energy of the states involved. The original Hohenberg–Kohn theorems (Hohenberg & Kohn, 1964) apply only to the ground state. The theorems may, however, be generalized to the energetically lowest state of any symmetry representation for which any property is a functional of the corresponding density. This allows (in cases where applicable) the calculation of excitation energies by taking total energy differences.

Many aspects of DFT from formalism to applications are discussed and many references are given in the book by Springborg (1997).

2.2.11. Band-theory methods

There are several methods for calculating the electronic structure of solids. They have advantages and disadvantages, different accuracies and computational requirements (speed or memory), and are based on different approximations. Some of these aspects have been discussed in Section 2.2.9. This is a rapidly changing field and thus only the basic concepts of a few approaches in current use are outlined below.

2.2.11.1. *LCAO (linear combination of atomic orbitals)*

For the description of crystalline wavefunctions (Bloch functions), one often starts with a simple concept of placing atomic orbitals (AOs) at each site in a crystal denoted by $|m\rangle$, from which one forms Bloch sums in order to have proper translational symmetry:

$$\chi_{\mathbf{k}}(\mathbf{r}) = \sum_m \exp(i\mathbf{k}\mathbf{T}_m)|m\rangle. \qquad (2.2.11.1)$$

Then Bloch functions can be constructed by taking a linear combination of such Bloch sums, where the linear-combination coefficients are determined by the variational principle in which a secular equation must be solved. The LCAO can be used in combination with both the Hartree–Fock method and DFT.

2.2.11.2. *TB (tight binding)*

A simple version of the LCAO is found by parameterizing the matrix elements $\langle m'|\mathbb{H}|m\rangle$ and $\langle m'|m\rangle$ in a way similar to the Hückel molecular orbital (HMO) method, where the only non-vanishing matrix elements are the on-site integrals and the nearest-neighbour interactions (hopping integrals). For a particular class of solids the parameters can be adjusted to fit experimental values. With these parameters, the electronic structures of rather complicated solids can be described and yield quite satisfactory results, but only for the class of materials for which such a parametrization is available. Chemical bonding and symmetry aspects can be well described with such schemes, as Hoffmann has illustrated in many applications (Hoffmann, 1988).

In more complicated situations, however, such a simple scheme fails.

2.2.11.3. *The pseudo-potential schemes*

In many respects, core electrons are unimportant for determining the stability, structure and low-energy response properties of crystals. It is a well established practice to modify the one-electron part of the Hamiltonian by replacing the bare nuclear attraction with a pseudo-potential (PP) operator, which allows us to restrict our calculation to the valence electrons. The PP operator must reproduce screened nuclear attractions, but must also account for the Pauli exclusion principle, which requires that valence orbitals are orthogonal to core ones. The PPs are not uniquely defined and thus one seeks to satisfy the following characteristics as well as possible:

(1) PP eigenvalues should coincide with the true (all-electron) ones;

(2) PP orbitals should resemble as closely as possible the all-electron orbitals in an external region as well as being smooth and nodeless in the core region;

(3) PP orbitals should be properly normalized;

(4) the functional form of the PP should allow the simplification of their use in computations;

(5) the PP should be transferable (independent of the system); and

(6) relativistic effects should be taken into account (especially for heavy elements); this concerns mainly the indirect relativistic effects (*e.g.* core contraction, Darwin *s*-shift), but not the spin–orbit coupling.

There are many versions of the PP method (norm-conserving, ultrasoft *etc*.) and the actual accuracy of a calculation is governed by which is used. For standard applications, PP techniques can be quite successful in solid-state calculations. However, there are cases that require higher accuracy, *e.g.* when core electrons are involved, as in high-pressure studies or electric field gradient calculations (see Section 2.2.15), where the polarization of the charge density close to the nucleus is crucial for describing the physical effects properly.

2.2.11.4. *APW (augmented plane wave) and LAPW methods*

The partition of space (*i.e.* the unit cell) between (non-overlapping) atomic spheres and an interstitial region (see Fig. 2.2.12.1) is used in several schemes, one of which is the augmented plane wave (APW) method, originally proposed by Slater (Slater, 1937) and described by Loucks (1967), and its linearized version (the LAPW method), which is chosen as the one representative method that is described in detail in Section 2.2.12.

The basis set is constructed using the muffin-tin approximation (MTA) for the potential [see the discussion below in connection with (2.2.12.5)]. In the interstitial region the wavefunction is well described by plane waves, but inside the spheres atomic-like functions are used which are matched continuously (at the sphere boundary) to each plane wave.

2.2.11.5. *KKR (Korringa–Kohn–Rostocker) method*

In the KKR scheme (Korringa, 1947; Kohn & Rostocker, 1954), the solution of the KS equations (2.2.10.3) uses a Green-function technique and solves a Lippman–Schwinger integral equation. The basic concepts come from a multiple scattering approach which is conceptually different but mathematically equivalent to the APW method. The building blocks are spherical waves which are products of spherical harmonics and spherical Hankel, Bessel and Neumann functions. Like plane waves, they solve the KS equations for a constant potential. Augmenting the spherical waves with numerical solutions inside the atomic spheres as in the APW method yields the KKR basis set. Compared with methods based on plane waves, spherical waves require fewer basis functions and thus smaller secular equations.

The radial functions in the APW and KKR methods are energy-dependent and so are the corresponding basis functions. This leads to a nonlinear eigenvalue problem that is computationally demanding. Andersen (1975) modelled the weak energy dependence by a Taylor expansion where only the first term is kept and thereby arrived at the so-called linear methods LMTO and LAPW.

2.2.11.6. *LMTO (linear combination of muffin-tin orbitals) method*

The LMTO method (Andersen, 1975; Skriver, 1984) is the linearized counterpart to the KKR method, in the same way as the LAPW method is the linearized counterpart to the APW method. This widely used method originally adopted the atomic sphere approximation (ASA) with overlapping atomic spheres in which the potential was assumed to be spherically symmetric. Although the ASA simplified the computation so that systems with many atoms could be studied, the accuracy was not high enough for application to certain questions in solid-state physics.

Following the ideas of Andersen, the augmented spherical wave (ASW) method was developed by Williams *et al.* (1979). The ASW method is quite similar to the LMTO scheme.

It should be noted that the MTA and the ASA are not really a restriction on the method. In particular, when employing the MTA only for the construction of the basis functions but including a generally shaped potential in the construction of the matrix elements, one arrives at a scheme of very high accuracy which allows, for instance, the evaluation of elastic properties. Methods using the unrestricted potential together with basis functions developed from the muffin-tin potential are called *full-potential* methods. Now for almost every method based on the MTA (or ASA) there exists a counterpart employing the full potential.

2.2.11.7. *CP (Car–Parrinello) method*

Conventional quantum-mechanical calculations are done using the Born–Oppenheimer approximation, in which one assumes (in most cases to a very good approximation) that the electrons are decoupled from the nuclear motion. Therefore the electronic structure is calculated for fixed atomic (nuclear) positions. Car & Parrinello (1985) suggested a new method in which they combined the motion of the nuclei (at finite temperature) with the electronic degrees of freedom. They started with a fictitious Lagrangian in which the wavefunctions follow a dynamics equation of motion. Therefore, the CP method combines the motion of the nuclei (following Newton's equation) with the electrons (described within DFT) into one formalism by solving equations of motion for both subsystems. This simplifies the computational effort and allows *ab initio* molecular dynamics calculations to be performed in which the forces acting on the atoms are calculated from the wavefunctions within DFT. The CP method has attracted much interest and is widely used, with a plane-wave basis, extended with pseudo-potentials and recently enhanced into an all-electron method using the projector augmented wave (PAW) method (Blöchl, 1994). Such CP schemes can also be used to find equilibrium structures and to explore the electronic structure.

2.2.11.8. *Order N schemes*

The various techniques outlined so far have one thing in common, namely the scaling. In a system containing N atoms the computational effort scales as N^3, since one must determine a number of orbitals that is proportional to N which requires diagonalization of $(kN) \times (kN)$ matrices, where the prefactor k depends on the basis set and the method used. In recent years

much work has been done to devise algorithms that vary linearly with N, at least for very large N (Ordejon *et al.*, 1995). First results are already available and look promising. When such schemes become generally available, it will be possible to study very large systems with relatively little computational effort. This interesting development could drastically change the accessibility of electronic structure results for large systems.

2.2.12. The linearized augmented plane wave method

The electronic structure of solids can be calculated with a variety of methods as described above (Section 2.2.11). One representative example is the (full-potential) linearized augmented plane wave (LAPW) method. The LAPW method is one among the most accurate schemes for solving the effective one-particle (the so-called Kohn–Sham) equations (2.2.10.3) and is based on DFT (Section 2.2.10) for the treatment of exchange and correlation.

The LAPW formalism is described in many references, starting with the pioneering work by Andersen (1975) and by Koelling & Arbman (1975), which led to the development and the description of the computer code *WIEN* (Blaha *et al.*, 1990; Schwarz & Blaha, 1996). An excellent book by Singh (1994) is highly recommended to the interested reader. Here only the basic ideas are summarized, while details are left to the articles and references therein.

In the LAPW method, the unit cell is partitioned into (non-overlapping) atomic spheres centred around the atomic sites (type I) and an interstitial region (II) as shown schematically in Fig. 2.2.12.1. For the construction of basis functions (and only for this purpose), the muffin-tin approximation (MTA) is used. In the MTA, the potential is assumed to be spherically symmetric within the atomic spheres but constant outside; in the former atomic-like functions and in the latter plane waves are used in order to adapt the basis set optimally to the problem. Specifically, the following basis sets are used in the two types of regions:

(1) Inside the atomic sphere t of radius R_t (region I), a linear combination of radial functions times spherical harmonics $Y_{\ell m}(\hat{\mathbf{r}})$ is used (we omit the index t when it is clear from the context):

$$\phi_{\mathbf{k}_n} = \sum_{\ell m}[A_{\ell m}u_\ell(r, E_\ell) + B_{\ell m}\dot{u}_\ell(r, E_\ell)]Y_{\ell m}(\hat{\mathbf{r}}), \qquad (2.2.12.1)$$

where $\hat{\mathbf{r}}$ represents the angles ϑ and φ of the polar coordinates. The radial functions $u_\ell(r, E)$ depend on the energy E. Within a certain energy range this energy dependance can be accounted for by using a linear combination of the solution $u_\ell(r, E_\ell)$ and its energy derivative $\dot{u}_\ell(r, E_\ell)$, both taken at the same energy E_ℓ (which is normally chosen at the centre of the band with the corresponding ℓ-like character). This is the linearization in the LAPW method. These two functions are obtained on a radial mesh inside the atomic sphere by numerical integration of the radial Schrödinger equation using the spherical part of the potential inside sphere t and choosing the solution that is regular at the origin $r = 0$. The coefficients $A_{\ell m}$ and $B_{\ell m}$ are chosen by matching conditions (see below).

(2) In the interstitial region (II), a plane-wave expansion (see the Sommerfeld model, Section 2.2.5) is used:

$$\phi_{\mathbf{k}_n} = (1/\sqrt{\Omega})\exp(i\mathbf{k}_n\mathbf{r}), \qquad (2.2.12.2)$$

where $\mathbf{k}_n = \mathbf{k} + \mathbf{K}_n$, $\mathbf{K}_n$ are vectors of the reciprocal lattice, $\mathbf{k}$ is the wavevector in the first Brillouin zone and Ω is the unit-cell volume [see (2.2.5.3)]. This corresponds to writing the periodic function $u_{\mathbf{k}}(\mathbf{r})$ (2.2.4.19) as a Fourier series and combining it with the Bloch function (2.2.4.18). Each plane wave (corresponding to $\mathbf{k}_n$) is augmented by an atomic-like function in every atomic sphere, where the coefficients $A_{\ell m}$ and $B_{\ell m}$ in (2.2.12.1) are chosen to match (in value and slope) the atomic solution with the corresponding plane-wave basis function of the interstitial region.

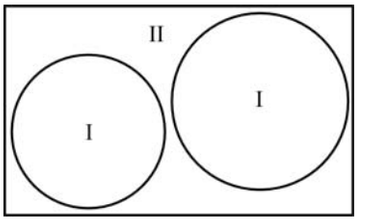

Fig. 2.2.12.1. Schematic partitioning of the unit cell into atomic spheres (I) and an interstitial region (II).

The solutions to the Kohn–Sham equations are expanded in this combined basis set of LAPWs,

$$\psi_k = \sum_n c_n\phi_{k_n}, \qquad (2.2.12.3)$$

where the coefficients c_n are determined by the Rayleigh–Ritz variational principle. The convergence of this basis set is controlled by the number of PWs, *i.e.* by the magnitude of the largest $\mathbf{K}$ vector in equation (2.2.12.3).

In order to improve upon the linearization (*i.e.* to increase the flexibility of the basis) and to make possible a consistent treatment of semi-core and valence states in one energy window (to ensure orthogonality), additional (k_n-independent) basis functions can be added. They are called 'local orbitals' (Singh, 1994) and consist of a linear combination of two radial functions at two different energies (*e.g.* at the 3*s* and 4*s* energy) and one energy derivative (at one of these energies):

$$\phi^{\mathrm{LO}}_{\ell m} = [A_{\ell m}u_\ell(r, E_{1,\ell}) + B_{\ell m}\dot{u}_\ell(r, E_{1,\ell}) + C_{\ell m}u_\ell(r, E_{2,\ell})]Y_{\ell m}(\hat{\mathbf{r}}). \qquad (2.2.12.4)$$

The coefficients $A_{\ell m}$, $B_{\ell m}$ and $C_{\ell m}$ are determined by the requirements that ϕ^{LO} should be normalized and has zero value and slope at the sphere boundary.

In its general form, the LAPW method expands the potential in the following form:

$$V(r) = \begin{cases} \sum_{LM} V_{LM}(r)K_{LM}(\hat{\mathbf{r}}) & \text{inside sphere} \\ \sum_K V_K \exp(iKr) & \text{outside sphere} \end{cases} \qquad (2.2.12.5)$$

where K_{LM} are the crystal harmonics compatible with the point-group symmetry of the corresponding atom represented in a local coordinate system (see Section 2.2.13). An analogous expression holds for the charge density. Thus no shape approximations are made, a procedure frequently called the 'full-potential LAPW' (FLAPW) method.

The muffin-tin approximation (MTA) used in early band calculations corresponds to retaining only the $L = 0$ and $M = 0$ component in the first expression of (2.2.12.5) and only the $K = 0$ component in the second. This (much older) procedure corresponds to taking the spherical average inside the spheres and the volume average in the interstitial region. The MTA was frequently used in the 1970s and works reasonable well in highly coordinated (metallic) systems such as face-centred-cubic (f.c.c.) metals. For covalently bonded solids, open or layered structures, however, the MTA is a poor approximation and leads to serious discrepancies with experiment. In all these cases a full-potential treatment is essential.

The choice of sphere radii is not very critical in full-potential calculations, in contrast to the MTA, where this choice may affect the results significantly. Furthermore, different radii would be found when one uses one of the two plausible criteria, namely based on the potential (maximum between two adjacent atoms) or the charge density (minimum between two adjacent atoms).